DOCUMENT RESUME

ED 067 394 TM 001 789

AUTHOR	Womer, Frank B.
TITLE	What Is National Assessment?
INSTITUTION	National Assessment of Educational Progress, Ann Arbor, Mich.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	70
GRANT	OEG-0-9-080771-2468(508)
NOTE	56p.
AVAILABLE FROM	National Assessment Staff Offices, Room 201A Huron Towers, 2222 Fuller Road, Ann Arbor, Michigan 48105 (single copies $2.00; orders of 10 or more, 20% discount)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Age Groups; Census Figures; *Data Collection; Elementary School Students; *Evaluation Techniques; Measurement Goals; *Measurement Instruments; *National Competency Tests; *National Surveys; Secondary School Students; Standards; Student Characteristics; Surveys; Young Adults
IDENTIFIERS	NAEP; *National Assessment of Educational Progress

ABSTRACT

National Assessment is a plan for a systematic, census-like survey of knowledges, skills, understandings, and attitudes designed to sample four age levels in ten different subject areas. It is an information-gathering program designed to provide both the educational community and the lay public with information about some of the direct outcomes of education as they are exhibited in students and young adults. The ten areas selected for assessment are Art, Career and Occupational Development, Citizenship, Literature, Mathematics, Music, Reading, Science, Social Studies, and Writing. Criteria of the National Assessment Committee in the setting of assessment of objectives include: (1) The objectives must be satisfactory goals for each subject area as seen by subject matter specialists; (2) The objectives must be ones which currently are accepted as goals of American education by most schools; and (3) The objectives must be ones which are acceptable to thoughtful lay adults as reasonable goals of American education. For related documents, see TM 001 793 and 797.) (Author/CK)

ED 067394

WHAT IS NATIONAL ASSESSMENT?

NAEP

ED 067394

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

WHAT IS NATIONAL ASSESSMENT?

Frank B. Womer
Staff Director

National Assessment of Educational Progress

Ann Arbor Offices:
Room 201A Huron Towers
2222 Fuller Road
Ann Arbor, Michigan 48105

Denver Offices:
822 Lincoln Tower
1860 Lincoln Street
Denver, Colorado 80203

A Project of the
Education Commission of the States

This publication was prepared pursuant to Grant No. OEG-0-9-080771-2468(508) with the Office of Education, U. S. Department of Health, Education, and Welfare. Grantees undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

National Assessment of Educational Progress, 1970
Library of Congress Catalog Card Number 77-115010

Single Copies $2.00
Orders of 10 or more, 20% discount

National Assessment Staff Offices
Room 201A Huron Towers
2222 Fuller Road
Ann Arbor, Michigan 48105

TABLE OF CONTENTS

Introduction 1

National Assessment Is An Idea 2

National Assessment Is Ten Areas 3

National Assessment Is A Set of Objectives 4

National Assessment Is A Set of Exercises 6

National Assessment Is A Set of Packages 18

National Assessment Is A Sampling Plan 20

National Assessment Is A Plan For Administration of Exercises 28

National Assessment Is A Scoring Plan 35

National Assessment Is A Reporting Plan 37

A Typical Report (hypothetical data) 41

National Assessment Is A Research Project 45

National Assessment Is An On-Going Project 46

National Assessment Is A Cooperative Project 48

National Assessment Is A Changing Project 50

What Is National Assessment? 52

Introduction[1]

Thousands of words have been written about National Assessment. Thousands more will be written. Many persons have written and spoken favorably about National Assessment; many others have been critical. National Assessment has been, and to some extent still is a controversial program. Too often both friends and foes project their own desires or their own fears for assessment or for evaluation into their conception of what National Assessment really is. National Assessment is a specific program. It does not encompass everything that has been written about it, either favorable or unfavorable. It is not the panacea for education that some hope; nor is it the evil monster that others fear.

What then is National Assessment? National Assessment is a plan for a systematic, census-like survey of knowledges, skills, understandings, and attitudes designed to sample four age levels in ten different subject areas. National Assessment is an information-gathering program designed to provide both the educational community and the lay public with information about some of the direct outcomes of education as they are exhibited by our students and young adults. The ultimate goal of National Assessment is to provide information that can be used to improve the educational process, to improve education at any and all levels where knowledge will be useful about what students know, what skills they have developed, or what their attitudes are. In brief, then, National Assessment aims at providing information in one of the many areas in education where more information is needed, the area of knowledges, skills, understandings, and attitudes.

Any definition or explanation of National Assessment, such as in the previous paragraph, summarizes succinctly the thoughts and ideas of those who are knowledgeable about the project already. Anyone hearing about the project for the first time would have to embellish those statements by conjuring up specifics of a plan that

[1]This monograph is adapted from the Second Annual Scates Memorial Lecture, delivered at the University of Florida, June 26, 1969.

would produce the information alluded to. Different persons would envisage quite different schemes. Yet National Assessment is a very specific plan that has been developed cooperatively by literally hundreds of individuals. While a relatively small group of men and women assumed the responsibility for developing the plans for National Assessment, they sought and used the advice of so many consultants and so many groups that no one person can claim that National Assessment is his creation.

In order to fully understand National Assessment one must understand the program in all of its many facets. This paper attempts to look at National Assessment from as many different angles as possible and to present the results of over five years of planning, as that planning has resulted in a specific plan for an assessment. Not every facet of the plan is unique; many parts have been adapted from other plans. As a sum total of its many parts, however, National Assessment is a unique educational project.

National Assessment Is An Idea

Any significant project begins as an idea, as an idea about something that needs to be done. In the case of National Assessment the idea began with thoughts of Francis Keppel, John Gardner, and others, as they speculated about the needs of education in this country and the criticisms that were leveled against education in the early 1960's. Defenders of education were and are hard pressed to come up with direct evidence that our schools are, in fact, doing a good job of meeting the needs of our society, as those needs are expressed in the objectives that schools set for themselves. To what extent are our 9-year-old students learning to read; to what extent are our 13-year-olds knowledgeable about the scientific aspects of our society, to what extent do our 17-year-olds understand the social structure of American life; to what extent have our young adults developed into thoughtful citizens? Neither these questions nor the hosts of others like them can be answered by information now available.

The "Idea" of National Assessment was to develop a plan for gathering direct information about knowledges, about understanding, about skills—information not currently available. One analogy to this idea is the development of the plan to gather systematically information about medical statistics, about incidence of disease and about other physical conditions of humans. Gathering of such

information proved to be a great spur to the development of better programs of public health. Discovering that the incidence of tuberculosis was considerably higher in low income families than in high income families, and that it was more prevalent in some parts of the country than others helped to guide the efforts at eradication. If National Assessment provides information that can be helpful in making wiser educational decisions, it will have achieved its goal, its "Idea." National Assessment is the "Idea" that accurate information about what boys and girls are learning is an essential ingredient for wise decision-making in education. Information alone does not change education; people who use information wisely can change education.

National Assessment Is Ten Areas

One of the earliest decisions that had to be made in the National Assessment Project was the choice of areas in which assessment should take place. A conservative approach would have been to select the 3-R's, and move ahead into the development of the plan. Instead, it was decided to include a much wider spectrum of subject areas. Education in the 1960's in these United States included reading, writing, and arithmetic; but education in this country in that decade also included art and music. To fully assess education one must include as many areas as are feasible. After considerable thought and review 10 areas were selected for National Assessment. These areas are Art, Career and Occupational Development, Citizenship, Literature, Mathematics, Music, Reading, Science, Social Studies, and Writing.

Several of these areas warrant a word of explanation. Career and Occupational Development represents a fairly recent change of title from the original one of Vocational Education. This change was made at the suggestion of a panel of representatives from the areas of vocational education, industrial arts, and vocational guidance. The objectives developed in this area are those of general education in occupational planning and in the development of generally useful skills, rather than being specific to the areas of vocational education or industrial arts.

Citizenship was developed as an area separate from Social Studies, an area that covers knowledges from the civics-political science areas and also includes a goodly portion of exercises covering citizenship behavior. It was felt that the area of

Citizenship was so crucial in our present society that it requires attention separate from the larger Social Studies area.

At the very beginning of the development of National Assessment it was anticipated that the project would be a growing and expanding one. Thus, it was anticipated that other areas would be added to the original ones. Consideration already is being given to the potential development of several areas in addition to the original 10.

National Assessment Is A Set of Objectives

As with any evaluative project, National Assessment is based upon a specific set of objectives. The objectives were developed by different agencies under contract to National Assessment. The methods they used varied a bit, but generally followed the same procedures. The literature was surveyed to see what other groups had done, and then subject matter specialists were brought together to evaluate, to expand, to elaborate, to edit, to give direction to the contractor. From the literature, from their own resources, and from consultants' suggestions the contractors produced the specific objectives of the Assessment for each of the 10 areas.

When the objectives were turned over to National Assessment an additional review of them was undertaken—a review by lay adults who were knowledgeable about education. Eleven panels of lay persons were organized. Each panel represented each of the four geographic areas and three types of communities—large cities, suburbs,[2] and small town-rural areas. Participants were selected from nominations made by the Congress of Parents and Teachers, the National Association of State Boards of Education, the National School Boards Association, and education committees of other organizations such as the National Association for the Advancement of Colored People and the U.S. Chamber of Commerce. Each of the 11 panels reviewed the objectives in all 10 areas. The chairmen of all panels then met together to consolidate the recommendations of all the panels. The result was that these lay panels accepted most of the objectives, but not all. They suggested editorial changes in a number of instances and for one subject area they asked for a complete revision. The original

[2]With one exception

objectives for the Social Studies area appeared to be lacking in clarity of direction for instrument development. The Social Studies objectives were reformulated in light of this suggestion.

The involvement of lay persons in the review of objectives is stressed here not because it was more important or more extensive than the reviews by subject matter specialists and other educators, but simply because it was a step that is not commonly undertaken by educators. It was an essential step for National Assessment, however, since the governing Committee had set three criteria for the development and acceptance of its objectives:

1. The objectives must be satisfactory goals for each subject area as seen by subject matter specialists.
2. The objectives must be ones which currently are accepted as goals of American education by most schools.
3. The objectives must be ones which are acceptable to thoughtful lay adults as reasonable goals of American education.

The various contractors, through their consultants, developed objectives that are deemed important by subject matter specialists and are accepted by the schools. The Committee's review by lay persons made sure that the third criterion was met.

The objectives developed for National Assessment will not satisfy everyone – no single set of objectives could. Since they were designed to reflect current practices and goals, they are not apt to cause many ripples of change. This will hearten some and disappoint others. It should be kept in mind that National Assessment is designed to be a candid camera looking at those knowledges and skills American youth have acquired; National Assessment is not designed to prod or push or coerce in any particular direction. It is designed to give educators and the lay public a clearer picture of where we stand. If a viewer does not like the educational picture he sees, he will be free to try to change that picture so that eventually it becomes a more pleasant view. While National Assessment is not designed to "change" American education, it is designed to provide information that lay persons and educators can use to effect change for the better.

The objectives developed for National Assessment will not become static. Already the objectives for the three areas to be covered in the first year of the Assessment are being reviewed and

revised, where necessary. From the beginning of the project it has been the intent to review the objectives of each subject matter area before that area is reassessed. National Assessment does not plan to be caught in a trap of assuming that once its objectives have been developed they can be used indefinitely. National Assessment itself may acquire expanded or even different goals over time, as it responds to ever-increasing demands for information pertinent to the evaluation of American education.

National Assessment Is A Set of Exercises[3]

The term "exercise" is used in the Assessment Program to cover all questions, items, or tasks that are used to gather information. By the time the reader finishes this section it should be clear why "question" or "item" are not appropriate terms to cover all of the specific tasks developed for National Assessment.

The initial sets of exercises were developed by the same contractors that developed the objectives for the 10 areas. Before they began their work they were given three directives:

1. to develop exercises in whatever form or mode seemed most appropriate to the assessment of a particular objective,
2. to develop exercises that were samples of some important knowledge, or skill or attitude, and
3. to develop exercises that sampled equally those attributes common to most assessees, to about half of those assessed, and to the ablest, most knowledgeable assessees of a given age.

Each of these criteria was so important that its accomplishment must be discussed in some detail.

Format or Mode of Exercises

Almost all of the questions used in standardized tests are multiple-choice exercises. Almost all standardized tests are designed to measure individual achievement or individual ability and to do it as reliably as possible by producing a rank order of

[3]An entire monograph on the subject of exercise development for National Assessment, by Carmen J. Finley and Frances S. Berdie, is in press.

individuals from low achievement to high achievement. In such a model the absolutely essential ingredient is to have questions that discriminate between those of lower and those of higher achievement. The multiple-choice exercise is ideally suited to that goal. Using recognition of an answer embedded in three or four foils (alternatives that might appear to be correct to a person not knowing the answer) has been demonstrated time and again to be a very efficient way to identify individual students of low achievement, individual students of average achievement, and individual students of high achievement. To identify high, average, and low performance for individual students one needs to provide a reliable rank order from high to low. It is not essential to know exactly what high achievers *know* or what average achievers *know*, but only that some assessees know more than others. National Assessment is not concerned with classifying individual students at all, but with identifying the things that defined groups of students can do. The model of standardized testing is *NOT* the model of National Assessment.

The classroom teacher, in the process of evaluating the performance of his own students, has somewhat different goals in mind than the goals of standardized testing. He is interested in determining the relative levels of performance of his students; but more than that he is interested in determining the level at which each student has mastered a given skill, or whether a student has acquired certain knowledges or understandings. Thus, a typical classroom teacher is more concerned with the extent to which his objectives of instruction have been attained by each student, than he is with whether Johnny knows more than Sue who knows more than Sam who knows more than. . . .

The classroom "mastery" model is not the National Assessment model either but it is much closer than the standardized test model. The major difference between the teacher's "mastery" model and National Assessment is that National Assessment is concerned with assessing the performance of groups of students (a sample of students) rather than assessing and reporting the performance of individual students.

This rather detailed attempt to differentiate between standardized tests and mastery tests is aimed at explaining the reason why National Assessment did not and does not feel constrained to limit its production of exercises to the type of question that has proven to be appropriate for standardized tests. National Assessment has

the freedom to use all of the techniques of measurement available to the classroom teacher plus a number of others that the teacher might not be able to use because of cost or practicality or limitations of time. This means, in practice, that the persons developing exercises for National Assessment were not limited to the use of the multiple-choice question. In fact they were urged to produce exercises that used whatever techniques that seemed most appropriate to a given task, the assessment of a specific objective.

With this directive in mind the exercise developers used by the various contractors did produce a fairly varied group of exercises, many of which are not in the multiple-choice format. Several examples might be noted. In the area of Science there are several exercises that are referred to as the "apparatus" exercises. They consist of various items of equipment which an assessee uses in a mini-experiment. An assessee actually uses the equipment to determine some experimental results which he then records. In the area of Writing almost all of the exercises require the assessee to write, not to check a box or oval. In the area of Citizenship there are many exercises that require the use of an interview technique and there are a few that are referred to as "group" exercises. In these, a group of eight assessees actually work together *as a group* on some question or problem. Their group performance is observed and categorized and will be reported. In the areas of Art and Music assessees will be asked to produce an "art" work[4] or to "sing along' (see Footnote 4) with a record or be recorded playing a musical instrument. It should be apparent now why the term "exercise" is preferred to the terms "question" or "item." Being asked to sing along with a record is not really a question.

While the number of types of exercises produced for National Assessment is considerable, a majority of the actual exercises produced are in fact, multiple-choice questions. This is unfortunate, not because the multiple-choice question may not be appropriate for the assessment of many objectives (in fact, it is), but because too many of the exercise writers seemed to take the easy way out by producing large numbers of multiple-choice questions and small numbers of direct measures of skills.[5] It is

[4]Performance will be judged (rated) by specialists.

[5]One exercise writer confided to the National Assessment staff director that he did not believe that National Assessment would really use direct measures of achievement because of expense and practicality, so he simply produced an entire set of multiple-choice questions.

difficult for many exercise writers to break the mold of conventionality, just as it is difficult for many others in education to do so.

As in most large projects, not all goals are achieved completely, particularly in the early stages of development. Thus, National Assessment will use many multiple-choice questions in its first cycle, along with a large number of open-end exercises and a smaller number of direct skill-type exercises. In future cycles it is anticipated that the ratio of direct skill measurement to indirect multiple-choice measurement will be reversed.

Content Validity of Exercises

National Assessment's one and only criterion of exercise validity is content validity. The most important directive that was given to the exercise developers was the directive to emphasize content validity above all other considerations. If an exercise has content validity it must be an exercise that is considered to be a direct measure of some important bit of knowledge or some important skill that reflects one or more of the objectives of a subject area. In practice an exercise has content validity if it "makes sense" to an informed reader who sees together an objective and an exercise designed to measure that objective; if the reader says "yes, exercise 102 is a good example of a skill called for in objective III-A."

The evaluation of content validity can vary, of course, depending upon the knowledges and skills of the person looking at an exercise. A mathematician can look at an exercise involving trigonometric knowledge and judge whether it "makes sense," whereas a lay person who never went beyond general math or beginning algebra may not know whether it makes sense or not. Thus, in technical areas, or in areas involving considerable knowledge, one must rely upon subject matter specialists to evaluate content validity. In less technical areas, such as the area of behaviors exhibited by a "good" citizen, any adult interested in the area may well consider himself a competent judge of the content validity of an exercise, and he may well be.

As an exercise developer sets out to produce exercises with content validity he must keep both the specialist and the lay adult in mind. Above all he must produce exercises that *are* meaningful, that *do* make sense, that *are* directly related to the objectives; that

are *not* trivia, that are *not* inconsequential, that are *not* peripheral to the objectives.

The exercise developers succeeded at distinctly different levels in producing exercises which do, in fact, have content validity. Some did a fine job, others not so fine. The National Assessment staff conducted a long series of review sessions—using subject matter specialists, using other professional educators, and using lay reviewers—to evaluate each and every exercise one-by-one. The major task given to these reviewers was to say either "yes, this exercise has content validity" or "no, this exercise does not have content validity." If the consensus was "no," the exercise was shelved. If the answer was "yes," further questions were asked, but generally it meant that that exercise remained in the pool of exercises, either as originally written or as later modified.

This review process means that every exercise used in National Assessment has been read and judged appropriate as a sample of a specific objective by a minimum of at least a dozen persons—subject matter specialists, other educators, lay persons, and the National Assessment staff and its regular consultants. But no group of from 12 to 20 persons can encompass all possible knowledge that conceivably could be brought to bear on an evaluation of content validity. Thus, even as the National Assessment progresses, questions will be raised as to the validity of some of the exercises. When National Assessment results are reported, the individual exercises will be subject to criticism by readers of the reports. Undoubtedly some of the criticism will point up problems and errors that were overlooked by the National Assessment reviewers. As the project grows and develops over the years, National Assessment will improve in general and its exercises will improve in content validity. Yet today, in 1970, the exercises prepared for National Assessment, with whatever faults they contain, may well be the best total set of exercises ever developed to attempt to tap directly those knowledges, skills, and understandings related to the major objectives of American education.

Difficulty Levels of the Exercises

In the development of questions for most standardized tests, item writers attempt to write questions so that about half of the examinees for which the test is aimed will be able to answer a

given question correctly. It can be demonstrated that a group of such exercises put together as a test does the best possible job statistically of producing a wide range of individual scores from low to high. That is, a test consisting of items of near 50 percent difficulty is the best way to reliably rank order individual students on the type of achievement or ability measured by the items.

But National Assessment is not designed to report scores on individuals. It is designed to provide meaningful information on the skills and knowledges that groups of individuals at a given age level have acquired. Thus the most appropriate difficulty level (or levels) for the exercises for National Assessment is not the same as for a standardized test.

One goal of the National Assessment is to report to the American public examples of knowledges, skills, and understandings that are common to *almost all* American youth. What level of reading skill has been developed by almost all 9-year-olds or by almost all 13-year-olds? What information about the world of science is common to almost all 17-year-olds? What understandings about the impact of American history on our current social structures are common to almost all young adults?

A second goal of National Assessment is to report to the American public examples of knowledges, skills, and understandings that are common to a *typical* or *average* American youth of a given age. What are some examples that are typical of the writing skill of 9-year-olds? What are some examples of the knowledges that a typical young adult has acquired about the world of music? What understandings does a typical 17-year-old have about the world of work?

A third goal of National Assessment is to report to the American public examples of the knowledges, skills, and understandings that only the *most able, most knowledgeable* American youth have acquired. What are some examples from the best artistic products developed by 9-year-olds? What are some examples of the level of knowledge acquired by 17-year-olds in trigonometry or advanced algebra? What are some examples of the understandings of literary interpretation that the ablest 13-year-olds have acquired?

In order to meet each of these goals of National Assessment—reporting knowledges or skills common to almost all, reporting skills or understandings of a typical student, reporting understandings or knowledges developed by the ablest—it was necessary to

direct the exercise developers to produce exercises and tasks at various difficulty levels. Specifically they were asked to aim at the following approximate goals:

1. One-third of the exercises at the 90 percent level (easy exercises)
2. One-third of the exercises at the 50 percent level (average exercises)
3. One-third of the exercises at the 10 percent level (difficult exercises)

No one expected that the exercise developers would or could produce a set of exercises that would meet these specifications exactly. Any person who could do that, while producing only exercises with content validity, could in fact produce the overall national results of National Assessment without doing the study.

The criterion of difficulty, then, was a guideline that was designed to insure the production of a large number of very easy exercises (90 percent or 95 or 85 or 80 percent could respond correctly) and a large number of very difficult exercises (only 5 or 10 or 15 or 20 percent could respond correctly) as well as the production of the more typical exercises (40 or 50 or 60 percent could respond correctly).

Establishing a criterion of exercise difficulty does not assure production of exercises that do meet this criterion. In this instance the exercise developers uniformly found it to be a very, very difficult task to produce very easy exercises, the ones referred to as the 90 percent exercises. In retrospect it is not too surprising; yet the severity of the problem was not foreseen. An initial tryout of a random sample of exercises labeled 90 percent by the exercise developers showed the exercises averaging in the 50 to 60 percent difficulty range. In other words the exercises designed to be so easy that 90 percent of all 9-year-olds could respond correctly were, in a trial situation, answered correctly by only 50 or 60 percent of 9-year-olds, on the average.

These results of initial tryouts led to the production of more exercises aimed specifically at the very easy end of the difficulty continuum. It was necessary to stress this criterion repeatedly while also stressing that content validity could *not* be sacrificed in the process. First an exercise must make sense; second it should meet the criterion of difficulty. Even though extensive field testing

was done on the exercises the actual results from the assessment will be needed to determine whether the criterion of producing exercises at three different difficulty levels was met as well as was hoped. If not, additional efforts will be necessary to meet this goal as National Assessment continues.

Directionality; Guessing; Understanding Exercises; Invasion of Privacy

In addition to the three criteria previously discussed, others were established to take into account various potential problems inherent in any large scale assessment. Several of them are important enough to warrant attention in any description of National Assessment.

Directionality: Each exercise developed for National Assessment was designed to assess at least one specific objective in a given subject area. Each objective spells out a specific goal that our society is striving to attain. Therefore each exercise, designed to sample whether an objective has been attained, must have a correct or best answer or a desired direction. If that were not the case an exercise would never provide any information about the attainment of an objective. This concept is referred to as "directionality." Each exercise developed for National Assessment must have directionality.

Directionality is easy to attain in the cognitive area. Each exercise does have a correct answer or best answer. Directionality becomes somewhat controversial in the affective domain, since many attitudes or interests do not have directionality, at least from the point of view of having any generally agreed upon correct or best answer. One could show a person reproductions of a Picasso, a Mattisse, a Braque, and a Van Gogh and ask him to select the "best" one. Responses to such an exercise would be impossible to "score" because there is no correct or best answer. Experts would not agree, let alone laymen. Such an exercise would not have directionality. One could report results as an indication of the art tastes of American youth, but it would provide no evidence as to whether any objective in art was being attained. On the other hand one could present an assessee with a reproduction of a Picasso along with reproductions of inferior imitators of Picasso and then assess whether a sample of assessees would select the Picasso when asked to select the best one. If art experts could

clearly agree that the Picasso was superior to the imitators, such an exercise would have directionality. In that case such an exercise could help to sample whether American youth are developing a taste for "good" art. Such an exercise would be in the affective area, but would have directionality. Some readers might protest that an assessee who happens to recognize a Picasso reproduction might answer for a cognitive reason rather than an affective reason. Such a situation could be handled by selecting an "unknown" Picasso or by using examples of buildings (architecture) or pots and pans (designs) and so on.

The point is, some survey-type questions of opinions do have directionality and some do not. All exercises included in the National Assessment, however, *must* have directionality. National Assessment is not a survey of interesting facts and attitudes, it is a census of knowledges, skills, understandings, and attitudes *that provide evidence as to whether important educational objectives are being attained.*

This latter phrase is of particular importance as the project proceeds to the point that one attempts to assess progress over time. One can assess progress only if movement in one direction is considered good. To find out that 40 percent of 17-year-olds in 1974 feel that the president of the United States should be elected for a six-year term of office whereas only 20 percent felt that way in 1969 might be interesting information to have, but it would not tell us anything about whether some objective in the area of citizenship was being met. On the other hand finding out that 90 percent of all 13-year-olds in 1974 would be happy to live in a neighborhood with people of many different religious beliefs whereas only 70 percent felt that way in 1969 could tell us that progress was being made toward an objective of greater tolerance toward persons with differing beliefs.

Directionality must remain an important ingredient of any plan to assess progress toward educational objectives.

Guessing: Since many of the exercises prepared for National Assessment are multiple-choice items, and since the goal of the Assessment is to report actual achievement levels, there is a problem of confusing actual knowledge with random guessing. It did not seem feasible or desirable to eliminate all multiple-choice exercises. The conventional "correction for guessing" did not seem appropriate since it does not in any way reduce guessing, but assumes that it can be averaged out over a series of questions.

From the point of view of actually reporting what assessees can do it would be ideal if everyone who might do random guessing on an exercise would refrain from making any response. In an ideal situation responses to an incorrect foil then could be reported as misinformation.

In order to come closer to the ideal than simply ignoring guessing and reporting that the percentage responses for each alternative is "somewhat" higher than they should be because of guessing, it was decided to attempt to discourage guessing by adding an additional alternative (foil) to most multiple-choice exercises. That alternative is "I don't know." A study was made to determine if assessees will actually use the alternative realistically. The study showed that use of the "I don't know" alternative did produce results that are closer to the results obtained from presenting the same exercise in a free-response format than from presenting it in the traditional multiple-choice format without "I don't know." The "I don't know" alternative is particularly useful with the very difficult exercises (10 percent) when one is certain that a large percentage of the assessees do not know the answer. In any case, it is intended that the exercises themselves will be presented in the final reports, so that each reader can see for himself the percentages that chose each alternative as well as "I don't know."

Understanding Exercises: In the early tryouts of samples of exercises from all 10 areas, some of the lower achieving examinees (identified by their own teachers) were interviewed individually after the tryout sessions. Each exercise was discussed with them in an effort to determine whether they truly understood each question or task that they were being asked to respond to. In too many instances it was discovered that lower achieving assessees were confused by a particular word or phrase, so that they had no opportunity to demonstrate whether in fact they had a certain skill or bit of information. In many instances a simpler vocabulary or a change in phrasing would solve the problem.

These early tryouts alerted both the staff and the exercise developers to the need to take a closer look at the exercises from the point of view of the assessee's understanding of each task or question, of simplifying language and phrasing as much as is humanly possible. During the discussion of this and related concerns a general principle of development and operation was enunciated. That principle was and is "do everything possible to

maximize assessee understanding of the tasks he is asked to perform." The simplicity of this principle might lull one to assume that it is applied to all testing, all assessment, all evaluation. In fact it is rarely applied in a truly rigorous fashion. In the National Assessment Program it has been used as a guideline both in exercise development and in the manner of presenting the exercises to assessees.

Invasion of Privacy: An issue of grave concern to many persons in our society is the continuing demand for information that is being made on all of us. Some of this accumulated information is fairly routine (date and place of birth, place of residence, occupation, etc.); other information is considered by many to be "not so routine" (credit ratings, personal beliefs, family relationships, etc.). The things that people are most sensitive to comprise an area with rather flexible, indeterminate boundaries. To attempt to refrain from gathering any information about every area that is potentially controversial would eliminate all data gathering. But there seem to be some areas of consensus about what invasion of privacy is. Some of them are written into existing legislation.

National Assessment, like any other data gathering project, could not ignore these concerns. Most of the National Assessment exercises are cognitive in nature and subject to very little concern by those most sensitive to the invasion of privacy issue. But even in the cognitive domain, it was necessary to consider exercises covering content which could be "touchy." Thus knowledge of human biology was included in science exercises; art exercises might include pictures of nude figures; and music and art could well include some religious works. A general principle was evolved that enabled National Assessment to minimize problems in this domain. The principle was "develop exercises which do assess objectives and which, if possible, avoid *unnecessary* controversy." The key word in that phrase is "unnecessary." If one can, for example, assess the objective of "Perceive and respond to aesthetic elements in art" without selecting reproductions of art works of voluptuous nudes it would be considered to be advantageous. The objective is assessed; no one is offended. There is no sub-objective in National Assessment that requires the assessment of voluptuous nudes in art. In contrast to that example it was felt that to eliminate exercises tapping knowledge about human biology would be an omission of one important knowledge objective in science. In that instance it was not possible to avoid a potential

area of concern because it would have meant failing to assess a sub-objective.

Of greater concern than the cognitive area for potential invasion of privacy was the affective area of opinions and attitudes. It is in this area that one finds the greatest diversity of opinion as to what is proper information to gather and what is not. In order to handle this concern all exercises written for National Assessment that could conceivably be considered to be an invasion of privacy (based on staff judgment) were reviewed by panels of lay persons. Not all exercises were taken to these panels, e.g., 2 + 2 = ?, etc. The lay panels discussed the various issues mentioned before and did make suggestions for replacing some exercises if an appropriate substitute could be developed that sampled the same objective satisfactorily. This proved to be a very manageable problem. In addition, the U.S. Office of Education, which is providing partial funding for the assessment, is responsible for examining all survey or test materials used in projects that it supports, for potential invasion of privacy. They asked that a few exercises either be revised or eliminated on the basis of potential invasion of privacy. This request was easily handled.

Because of an emphasis upon objectives rather than content it was possible to substitute non-offensive exercises for potentially offensive ones in most instances. Nevertheless, a few exercises were retained which, it was realized, might be offensive to some minority of the population. Because of this, and because of the completely voluntary nature of the assessment, it was decided that *any school selected in the sample could eliminate any specific exercise or exercises for its students,* and that any assessee could refuse to answer any exercise he chose to.[6] As an additional protection for the individual assessee, his name was not recorded on the booklet of exercises. The records of names and serial numbers of assessees were never removed from a school building, and will be destroyed by the principal after a reasonable period of time during which they are available to recover lost information, e.g., age or sex of an assessee, through the building principal only.

Experience with the first phase of the assessment demonstrated that the staff had not foreseen every possible problem. One state law restricts the use of any questions relating in any fashion to attitudes about family life, morality, or religion without parental

[6]Very few schools or individuals exercised this option in the first year.

permission. Exercises asking about attitudes toward persons of a different race had to be omitted in that state on the basis of a legal opinion there. In these instances the resulting statistics will have to be reported on a reduced sample size, and thus, will be less reliable than if no loss had occurred.

Since the question of invasion of privacy is a highly personal one, it is doubtful if any project of the scope of National Assessment could possibly anticipate every potential concern that might arise. Feedback from school personnel and parents themselves should, however, bring the project close to the point of eliminating this concern.

National Assessment Is A Set of Packages

Most persons refer to the National Assessment "tests"; the staff refers to the National Assessment "packages." To some persons this is merely a play on words. To the staff it is a deliberate choice of a term, "package," that is meant to point up the fact that National Assessment has put its exercises together in a fashion that is different from the method generally used in developing tests. The greatest difference between a National Assessment package and a test is that almost all of our packages contain exercises from two or three different subject areas, whereas a test contains exercises from a single subject area.[7]

A test is designed to yield a meaningful score for each individual student who takes it. Thus it must consist of questions that "hold together," that make sense when the results are cumulated or scored in any fashion. This is not true in National Assessment. Package number 1 for use with age 17 assessees in assessment year 01 contained 11 exercises. Of these 11 there were seven multiple-choice Science exercises, three free-response Citizenship exercises, and one essay Writing exercise. If one attempted to add scores from seven Science exercises plus three Citizenship exercises plus one Writing exercise the total score would have no meaning. But the purpose of National Assessment is to report separately for each exercise, not to report a score for an individual assessee. Therefore the project was free to package the exercises in any convenient fashion that added up to about 40 or 45 or 50 minutes of assessment time for each assessee. Because of this difference,

[7]A test "battery" may cover several subject areas.

the NAEP staff will continue to work and talk about the National Assessment "packages." Many others, no doubt, will continue to talk about the National Assessment "tests."

Several additional points need to be made about the National Assessment packages. Most of the exercises are of the type that can be administered in a group situation (a maximum of 12 assessees in a group in assessment year 01), but some must be administered individually. Therefore, a few packages consist entirely of exercises that must be administered individually.

In addition to varying subject matter within each package, exercises of all three difficulty levels were included in each package. An "easy" exercise was used to start each package, and thereafter there was an alternation between exercises of easy, average, and difficult levels. The exercises were *not* scaled in the sense of putting all easy ones first and all difficult ones last. It was felt that that method would be more discouraging to assessees than the method of alternation and would cause some to stop working prematurely.

Generally the multiple-choice exercises were placed first in each package, followed by short-answer free-response exercises, and with the longer essay type exercises coming last. This was to simplify administration of the packages, so that time could be called if all assessees completed a final "20-minute" Writing exercise in less than 20 minutes. In addition it made it simpler to use separate sample exercises for each section.

An attempt was made to produce attractive packages—a larger than normal type face was used and a single exercise was placed on each page. Thus, physically the National Assessment packages don't look like tests. Since there was little external motivation for assessees to do their best, packages had to be appealing.

There are about 12 packages for each age level, each one containing different exercises. Each assessee takes one package only.

An interesting side light to the development of the packages has been the reaction of the assessees. Typical comments have been that "it's not like other tests," "it's fun," "I wish we had more tests like that," "it was hard," "it was easy." By and large assessees seem to enjoy taking a National Assessment package.

National Assessment Is A Sampling Plan

The National Assessment sampling plan has three aspects to it:

1. Sampling of exercises from the universe of all possible exercises
2. Sampling of four age groups from all possible ages
3. Sampling of individual assessees from the defined populations (age-group universes)

These aspects of the National Assessment sampling plan are not unique, individually, but the way they have been combined in National Assessment is unique.

Sampling of Exercises

In any given subject area an almost infinite number of exercises could be prepared. One could never even think of being exhaustive or even semi-exhaustive in covering a subject area. The preparation and use of a limited number of exercises is inevitable. The representativeness of the exercises selected to be used in such a project is crucial—they must be spread across all major objectives and must attempt to be representative of varying content areas within a subject field as well. Thus, in Science one not only wants exercises that sample scientific knowledge, but one wants such exercises to include chemical as well as physical areas, biological as well as geological areas, and so on. While it would be nice to be able to sample each specific sub-content area for each objective, the total number of exercises necessary for such a task would be much too great for any project without limitless funds. Therefore one must compromise, using as many exercises as possible while being very careful to spread the ones used across objectives and content areas.

Time available for administration became the crucial variable in setting the upper limits for the number of exercises to be used during the first year of the assessment. On the average about 160 minutes of assessment time was available for each subject area for each age. This was dictated by monies available for the actual field administration work. This meant that for each age group selected, all of Science had to be sampled in about 160 minutes, all of Writing in about 160 minutes, and all of Citizenship in about 160 minutes. When multiple-choice exercises were used primarily, as in

Science, quite a few could be asked in 160 minutes. When essay exercises were used primarily, as in Writing, only a few could be asked in 160 minutes.

This constraint upon the total time available to a given area means that the exercises used represent only a minute portion of the multitudes that could have been used. The exercises used, then, are a sample only, but were carefully chosen to be representative of the total universe of exercises. When they are reported they must be interpreted as a sample of the exercises, not as the domain of all important knowledges, skills, understandings, and attitudes in a given subject area.

Sampling of Age Groups

The population of assessees for National Assessment is defined as all 9-year-olds, all 13-year-olds, all 17-year-olds, and all young adults ages 26 through 35 in the 50 states plus the District of Columbia. The only exception to this definition is the exclusion of institutionalized individuals of these given ages, those in hospitals, in prisons, etc., who could not be reached except by extraordinary means. The percentages of such persons is very small compared to the total population.

The choices of age groups were made to provide information at particularly meaningful periods in the educational life of Americans. At age 9 most students have completed their primary education; at age 13 most students have completed their elementary education; and at age 17 most students are close to the end of their secondary education. The use of these three ages provides a uniform four-year age differential, which was deemed desirable. The adult age group was defined as individuals between the ages of 26 and 35, a 10-year span. The age of 26 was chosen as an age at which most adults have completed all of their formal educational work. A 10-year span was chosen to provide a very large population from which to sample, and to avoid any vagaries of educational practices that might have affected individuals of a given single-year age group.

The choice of age groups rather than grade groups was made to provide a more uniform, meaningful, understandable category. Because of differences across the country in age-grade placements and because of differences in promotion policies it was felt that grade groups would be more diverse (less homogeneous) in their

educational attainments than would age groups. Further, it was felt that most laymen (non-professional educators) would be able to better understand and interpret results by age groups than by grade groups.

Sampling of Individuals

Once a population is defined it becomes necessary to consider how to select a random probability sample that is truly representative of that population. It would be patently impossible to assess three or four million 17-year-olds, for example. It has long been known that a very small sample is entirely adequate for gathering information, provided one secures the cooperation of a high percentage of the individuals selected in that sample. National polls, such as Gallup, Roper, Harris, use only a small fraction of 1 percent of their defined population to predict election results and other attitudes of the American electorate. The National Assessment sample for each exercise is comparable in scope to those of the national opinion polls.

Ideally it would be nice to have the name of each and every 9-year-old, put the names in a hat, and draw a random sample. Such a procedure is not at all feasible for a research study such as National Assessment–the task of collecting names would be gigantic, the location and administration of exercises to such a sample would be stupendous.

For ages 9 and 13 it was decided to use a school sample only and for ages 26 - 35 a household sample only. For age 17 it was decided to use both a school sample (since the majority of 17-year-olds are in school) and a household sample (since a sizeable minority of 17-year-olds are not enrolled in any school).

For the school age sample what one does is to select a sample of small geographic areas, such as counties and cities, then locate and sample school buildings in the geographic areas, and finally sample the youngsters of a given age within the buildings selected. This procedure will be elaborated on a bit, but no attempt will be made here to present the sampling plan in a precise statistical manner. Other, highly technical, publications are available for those interested in the details.

An important principle to keep in mind is that each person in the defined population (each 17-year-old, for example) has an equal or at least known chance (probability) of being selected in

the actual sample. In the final computation of results the *known* probabilities are used to weight the sample results so as to give unbiased estimates for the populations sampled. For most nonstatisticians it is satisfactory to think of a sample being chosen so that everyone has an equal chance of being represented.

No sample, no matter how drawn, can guarantee absolute accuracy. All results must be interpreted within the potential error commonly called sampling error. This potential error can and will be computed and reported with the final results themselves.

The actual first step in developing the National Assessment sample involved dividing the entire country into geographic units, as follows:

1. cities
2. counties (exclusive of the cities)
3. pseudocounties (two or more counties put together when the population of a single county was less than 16,000)[8]

Each city, county, and pseudocounty was assigned a unique number for *each* unit of 16,000 persons residing therein. Thus a city of 16,000 population was assigned one number, a city of 32,000 got two numbers, a city of 160,000 got 10 numbers, a city of 1,600,000 got 100 numbers, etc. Thus every 17-year-old (or 9- or 13-year-old) was included in one potential sample unit, with a population of about 16,000.

Then, in order to insure comparable representation in the final sample for each part of the country, four geographic areas were developed (Northeast, Southeast, Central, and West). An equal number of sampling units was chosen from each geographic area—52 from each, or 208 nationally.

And before the actual selection of the sampling units, each geographic region was divided into communities of four types, e.g.,

1. large cities (above 200,000 population)
2. urban fringe (communities adjacent to the large cities)
3. middle size cities (25,000 to 200,000)
4. small town-rural (below 25,000)

[8]The number 16,000 was selected as a population large enough to guarantee sufficient 9-, 13-, and 17-year-olds to secure 150 assesses at each age level after all possible losses; 16,000 is more than would have been needed to guarantee sufficient adults within the ages of 26-35 for the sample.

The 52 sampling units for each geographic area are spread across the four community types in a fashion proportional to their population in relation to the total area population. For example, in the Northeast 15 of the 52 sampling units were chosen from large cities, whereas in the Southeast only eight of the 52 sampling units were chosen from large cities. In contrast, the Southeast has 23 sampling units from small town-rural areas, while the Northeast has seven in this category. This is because there are more people in the Northeast living in large cities than in other types of communities, whereas in the Southeast more people live in a small town-rural area.

Within each geographic region and within each community-type the actual sampling units chosen were selected at random, by use of a technique that uses tables of random numbers. For example, the 15 large city sampling units in the Northeast were chosen at random from all cities in the Northeast with a population above 200,000. Many of these cities did not actually fall in the random 15 sampling units selected, and by chance several of the very large cities had more than one sampling unit in the final sample.

This scheme did *not* guarantee that all large cities be included in the sample; it only guaranteed that a sample of large cities in the Northeast, in the Southeast, in the Central region, and in the West would be included. This scheme guaranteed that a sample of all urban-fringe cities would be included in each of the four areas, that a sample of all middle-size cities would be included, and that a sample of all small town-rural areas in each geographic area would be included. This scheme did *not* guarantee that all 50 states be included in the sample, only that all 50 states had a chance of being included. In fact, only 39 states fell in the sample.

It is extremely important to keep the sampling plan in mind when considering reporting possibilities. It would not be statistically sound to report results for units that were not sampled adequately by the plan. Obviously, complete state results could not be reported because not all states were included in the sample. And in those states that contained only a very few sampling units, results would be so variable and based on such a small sample as to be meaningless.

However, it is sound statistically to report for some breakdowns that were not built into the sampling plan, e.g., by sex. The sampling plan did not call for sampling boys and girls separately, but experience (and common sense) tells one that both boys and

girls will be included in a sample as large as 25,000 or 30,000, without having to plan for it specifically.

After the selection of the 208 sampling units (52 in each geographic area) it was necessary to take several more steps. Each sampling unit contained about 16,000 persons. National Assessment required a random sample of all 17-year-olds, 13-year-olds, 9-year-olds, and adults from 26 - 35 in that population. It was decided that the best way to find school-age assessees was through schools. So, for each age, all school buildings enrolling students of that age, both public and private, were identified by various available source books or other references, along with approximate building enrollments.

The actual next steps that were taken were very detailed statistically. They were designed to produce units of approximately 150 pupils from at least two different buildings within each sampling unit. For the first age 17 assessment (March, April, May of 1969) 673 actual school buildings were selected to represent the 208 sampling units.

The next step was a practical one, securing the cooperation of each school selected. This was done through initial contacts with each school superintendent and then each building principal. Approximately 87 percent of the schools in the 208 sampling units for age 17 agreed to participate. For the first age 13 assessment (October and November of 1969) 96 percent of the schools agreed to cooperate and for age 9 (December, 1969 through February, 1970) 95 percent agreed. This is an exceptionally high percentage of cooperation, particularly considering the original controversy over the value of the project itself. This high percent of acceptance guarantees that the school age results will be highly reliable.

The final stage of sampling took place in the schools themselves. It was necessary to ask each building principal to provide a list of names of all students of a given age, say age 17. This listing (which never left the school building) was used for the final random selection of actual boys and girls to take the assessment exercises. For example, assume the following figures:

Sampling unit number 159	Enrollment of 17-year-olds	Number in Sample
School A	300	75
School B	200	50
School C	100	25

The selection of the 75 assessees for School A or the 50 for School B or the 25 for School C was done by the use of random numbers and selecting from the totals of 300 or 200 or 100. In practice, "extra" students were selected to cover potential absentees on the assessment days. For example, about 100 were selected for School A, 25 being alternates, and so on.

This then is the National Assessment in-school sampling plan for ages 9, 13, and 17 for identifying individual assessees–with *one major exception.* The exception took place during the second and third parts of the sampling. Its purpose was to increase the accuracy (reliability) of the results from one segment of the population without, statistically, disturbing the other results. A very simple example has been prepared just to illustrate the process.

Suppose that one wanted to be sure to get a reliable sample of left-handed assessees and *suppose* that it is known that only about 20 percent of the population of all assessees are left-handed. If one were to use regular sampling techniques described above, one would expect to get, from a sample of 30,000 assessees, only 6,000 left-handed assessees and 24,000 right-handed assessees. Results from 24,000 assessees are bound to be more reliable than results from 6,000. Thus the results for left-handed assessees would not be as accurate as for right-handed assessees. What to do?

Suppose that it were possible to identify lefties prior to final sample selection so that one could sample, randomly, 15,000 left-handed persons and 15,000 right-handed persons. In this case the results would be equally reliable for both types of assessees. But wait, since one also wants to report total results accurately, the combination of 24,000 right-handers and 6,000 left-handers would be appropriate whereas the combination of 15,000 plus 15,000 would *not* be appropriate. Statistically, it is very simple to use the 15,000 and 15,000 breakdown and still come up with unbiased total results. It could be done simply by counting each of 15,000 *right-handers'* scores four times each. That would yield 60,000 right-handers' scores to be added to 15,000 left-handers' scores–exactly the same ratio as the 24,000 to 6,000, or 80 percent to 20 percent.

Now, be assured that right-handers and left-handers *were not* sampled for National Assessment, but low SES and high SES students were sampled in a fashion much like the one described.

SES is a common abbreviation for socio-economic status. In National Assessment it is an abbreviation for socio-educational status. It is an index that attempts to classify assessees on the basis of being privileged or under-privileged, from the point of view of both sociological and educational opportunities available to an individual. National estimates suggest that approximately 15 percent to 20 percent of the total population could be considered to be disadvantaged. It was felt that to be able to report really reliable results for low SES assessees one must oversample them, in a fashion much like the one described for left-handers. This was done. It was done on the basis of choosing a higher than "normal" percentage of school buildings that were located in low income areas. It was very simple to get information about areas of low income from census data and to choose extra schools in these areas. Thus, the National Assessment sample was designed to oversample low SES assessees. But, in order *not* to bias total assessment results the low SES assessees were used in the total statistics at exactly the same percentage rate that they exist in the total population. Total statistics were *not* overweighted with low SES results, yet the low SES results alone are more reliable than they would have been without using this technique.

Use of this sampling technique created some concerns on the part of school superintendents and principals. Because of an excess of low SES schools it appeared on the surface that the National Assessment sample was not truly random. This was of particular concern to the superintendents of a few school districts in which, by chance, the two or three or four schools selected happened all to be low SES, central city type schools. What was not readily apparent to them was that another city had high SES schools. Under the restrictions placed upon this sample it was inevitable that some systems would have only low SES type schools and some only high SES type. The built-in oversampling of low SES schools accentuated this phenomenon. Statistically, the method used was and is sound; politically it is difficult to explain. The left-hand right-hand example is an attempt to make it meaningful.

As was stated earlier, this explanation of the National Assessment sample is a simplified one. There were various minor variations from the general plan. All of them were statistically sound. The basic principle remains—the sample was and is representative of the defined populations.

Out-of-School Sampling

All of the foregoing explanation is for the in-school sample. The adult sample was drawn from households rather than from schools. The first part of the procedure, selection of 208 sampling units, was identical for both the school and household samples. They differ only for the second and third stages. The 208 sampling units were geographic areas containing about 16,000 persons. The goal was to randomly select 100 adults, ages 26 to 35, and get them to answer a sample of assessment exercises. In order to do this simply, each geographic area was divided, on the average, into about 160 blocks or other relatively small areas called segments. Ten of these were then selected at random from the 160 thus constructed. Within each area segment the National Assessment interviewers knocked on doors of designated homes and apartments and asked if anyone between the ages of 26 and 35 lived there. When the answer was yes, an appointment was requested for the actual assessment. Sometimes it could be done at once; more often it required a call back. This procedure is much simpler to explain than the in-school sampling, but it is a much more expensive procedure to operate because one gets only zero or one or two or three persons per household whereas one may get as many as 75 to 100 assessees in a school building. Again the key to success (reliability) of the sample is the percentage of acceptance.

The adult household sample was also used to secure a sample of out-of-school 17-year-olds. Not all 17-year-olds are enrolled in school, so a school sample of 17's will miss drop-outs and 17-year-old high school graduates. It was felt that the number of out-of-school 17-year-olds is large enough that an attempt had to be made to secure a sample for the assessment. In each household contacted for the adult assessment, an inquiry was made as to whether a 17-year-old lived there. If the answer was yes, the next question was whether the 17-year-old had been enrolled in high school the previous April 1. If no, that 17-year-old was eligible for the National Assessment sample and was asked to participate in the assessment.

National Assessment Is A Plan For Administration of Exercises

There are two commonly understood plans for the administration of tests in national testing programs. One of these is the

scheme used for "secure" testing programs like College Boards, American College Testing Programs, etc. Basically, those programs hire administrators within schools and colleges, who devote a few half days a year to setting up, organizing, and administering the tests in a local community. Test materials are mailed to them and they return the resulting marked tests to the agency that prepared the materials. Examinees are self-selected generally, those who wish to apply for college or want to take the tests for some other reason.

The second common plan for administration of tests nationally is that used in standardizing or norming tests. The general procedure for norming is for a publisher first to secure cooperation of a set of schools (almost never really randomly selected), to ask a counselor or teacher or administrator in that school to administer the tests to a certain number of classrooms at a given level, to send the materials to the designated local administrator, and to receive them back after the administration.

In both patterns an attempt is made to secure persons with some experience in test administration, to administer the tests on a one-shot basis. Administrative procedures are developed carefully by the publisher, and all local administrators are urged to follow them exactly.

The National Assessment plan for administration of its packages (tests) has some similarities with existing patterns, but it also has some distinct differences.

Trained Staff

One important difference is that a full-time trained staff of 27 administrators has been employed by National Assessment[9] to handle the field work. These 27 individuals are called district supervisors (DSs) and are located all over the country, each one serving a specific geographic area. They have direct responsibility for contacting schools, for securing and training local persons to help in administration, and for making all arrangements in their own area. These district supervisors meet together as a group three times a year for orientation, training, sharing of ideas, receipt of new materials, etc. They also maintain day-to-day contact with

[9]Through its two contractors for administration, Research Triangle Institute (RTI) and Measurement Research Center (MRC).

either RTI or MRC in relation to specific questions and concerns that arise. They submit weekly reports of their activities.

One of the principal tasks of each DS is the hiring and training of local exercise administrators (EAs). The EAs administer most of the group administrable exercises in the schools and do most of the household interviewing. The DSs themselves handle certain specific administrations, often the individual ones in the schools. EAs are hired on an hourly basis. Sometimes they work within a single sampling unit (in geographically separated areas); sometimes they work within several sampling units (in geographically clustered areas). EAs are recruited from lists of substitute teachers, from among college trained housewives, from graduate students (summer assessment), and from other sources of competent adults. Each EA is trained for his task by the DS in charge in a given area.

The aspect of this plan that differs from national testing programs is that National Assessment is using its own staff of administrators, with training in handling National Assessment materials.

Age Groups

Since National Assessment is designed to sample people by ages rather than by grades, the administration of the exercises in the schools is different from other programs. Intact classes cannot be used; assessees are drawn from different classes and must be brought together in one room for the administration. This puts something of a burden on the schools cooperating in the assessment. However, it has not proven to be a serious problem. School schedules are flexible enough that very few concerns have been expressed with this aspect of National Assessment.

Taped Administration-Group Administrable Packages

A distinctly different aspect of the administration of National Assessment packages is the use of taped directions and taped reading of the exercises. A basic principle of National Assessment is to maximize understanding of the task to be done or the question asked. There was considerable evidence in the early tryouts of exercises that many assessees, particularly low-achieving

assessees, often were unable to respond to a particular task or question simply because they did not understand a word or phrase essential to an understanding of what they were being asked to do. In many of these instances an assessee could respond to the exercise when the troublesome word or phrase was explained or defined in terms familiar to him. One result of this phenomenon was an attempt to simplify language.

Another result was the use of taped reading of all group administrable exercises as well as of the general directions for administration. The one exception to this practice will be when the area of reading itself is assessed.

Specific research was undertaken to discover whether the use of taped reading of exercises would produce any different results than those when tape wasn't used. The research indicated that for assessees who are poor readers and for assessees from bilingual homes, the use of taped reading of the exercises did increase group performance. At the lower age levels the increase in ability to perform the exercises was greater than 25 percent among low achievers. At the same time the use of tape did not inhibit the performance of average and above average readers.

A problem allied to the use of taped administration is the choice of voice or voices to use. Research in several different regions indicated that the use of a television or radio announcer was just as effective as the use of a local teacher speaking with a regional accent. For this reason a professional announcer was used to record the administrations on tape.

Individually Administered Packages

Several packages at ages 9, 13, and 17 consisted of exercises that had to be administered individually. These were administered in an interview type situation, with only the assessee and the DS or EA present.

Adult Interviews

The administration of the packages for the adult assessment was done in an interview arrangement. Certain exercises were read aloud to the assessee and his responses were recorded by the interviewer. The multiple-choice and Writing exercises, however, were presented by handing a booklet to the adult and asking him

to complete them by filling in an oval beside his choice of an answer or by writing out his response–just as was done for the group administrable exercises for the in-school sample.

Length of Package

Each package was designed to require about 50 minutes of administration time. Timing for the group packages was exact since it was done on tape and, therefore, was paced for each exercise. Each assessee took only one package, with the exception of the out-of-school 17-year-olds who were asked to take four packages each. This was because the percentage of out-of-school 17s is so small that it would have been prohibitively expensive to go to sufficient households to find an adequate number of the sample. Each out-of-school 17-year-old was paid $10 for his time.

Anonymity

An assessee was never asked to record his or her name. A code number on the booklet was necessary in order to be able to identify each booklet as to sex, age, geographic area, etc. The only connection between that code number and an assessee's name was a roster kept by the school and never removed from the building. The roster was necessary in order to draw the in-school sample randomly from among all those of a given age enrolled in that school.

School Contacts

National Assessment, like other projects, is dependent upon the voluntary agreement of schools and individuals to participate in the project. Thus, it was essential to develop a plan to inform all interested and concerned parties in the process of asking school permission for ages 9, 13, and 17. The procedure for securing acceptance in the households is simpler, but cooperation is much more difficult to obtain.

Individual school contacts were preceded by a general letter to each governor, each chief state school officer, and the executive officer of each state branch of the American Association of School Administrators, National Association of Secondary School Principals, and the Department of Elementary School Principals, where

they are organized. These letters outlined the project and asked for general support of National Assessment.

The next level of contact was with the school district, through the superintendent of schools. A general letter was sent, followed by a phone call asking for an appointment to discuss school district cooperation in the National Assessment Program. In some instances the superintendent made the decision to participate or not participate. In other situations members of his staff, including the principals of the buildings selected, were asked to participate in the decision.

The degree of school district acceptance for the very first phase, age 17 in March through May of 1969, was very encouraging, nearly 90 percent. For ages 9 and 13 in the first assessment year it was above 95 percent.

Each participant had the right to refuse to cooperate if he chose, or to not answer any exercises that he objected to. This option was rarely exercised. Most assessees seemed to enjoy the opportunity to take a "test" that didn't seem like a test.

School Involvement

The schools that participate in National Assessment are asked to do three things:

1. Provide a room or rooms for the assessment to take place
2. Provide a listing of all eligible assessees in the building (all 17-year-olds or all 13-year-olds or all 9-year-olds)
3. Provide a building coordinator to help arrange the scheduling and movement of assessees from classroom to assessment room

Schools generally had very few complaints about the procedures or the requests made of them. A few felt that they needed more lead time to prepare lists or arrange for the scheduling on the assessment days. And in a few instances specific problems arose related to a local situation. The total operation at age 17 during the first year was very, very smooth.

Assessment Week

A week was allocated for assessment in each sampling unit. On

Monday the District Supervisor chose each building sample from the list of all eligible assessees of the given age. He developed the specific schedule for a building with the local coordinator, and he trained the Exercise Administrators for that building. Tuesday, Wednesday, and Thursday were the assessment days, and Friday was a "clean-up" day to be used if absenteeism or some other reason necessitated it. This was the general pattern, although there were quite a few variations that were developed to meet local needs.

Household Administration

The household assessment followed accepted procedures for interviewing adults in the home. After sample households had been selected, an interviewer made up to four personal contacts in an attempt to find someone at home. When a home was contacted the person available was asked for the name of anyone living in that household who was between the ages of 26 and 35 or who was age 17[10] and had not been in school the previous April 1. Then those persons who fell in the appropriate age ranges were asked to participate in an educational survey, right then, or to make an appointment for the actual interview. Many times an appointment had to be made since the person of the appropriate age was not at home at the time of the initial household contact.

The percentage of cooperation of adults and out-of-school 17s was not nearly as good as for in-school assessees. During the first assessment year the percent of cooperation *of those individuals contacted* was about 60 percent. It must be kept in mind also that other individuals were not included because, in some households in the sample, no one was ever found at home after four calls (including evening and weekend calls). Sixty percent is not nearly as good as the 90 percent school acceptance and is below the percentages generally experienced by survey organizations.

A special quality check of the first summer assessment is underway. An attempt is being made to secure the cooperation of a portion of non-respondents, to judge whether any significant

[10]Out-of-school 17s were defined as individuals between 16½ and 18½ years of age. This was done to enlarge the potential sample. It was felt that most out-of-school 17s would be drop-outs, whose achievement levels on National Assessment exercises would not vary too much over a two-year span.

bias has been introduced into the results. Most people assume that non-respondents are different from respondents in their survey responses to attitudinal questions, but we need to know in fact if and to what extent they are different on an achievement survey.

National Assessment Is A Scoring Plan

In a very real sense the word "scoring" is not appropriate for the process of summarizing the results of each National Assessment exercise. The goal is to be able to report, to summarize, in as meaningful a fashion as possible, the behavior exhibited by groups of representative individuals. An exercise "score" (correct or incorrect) is but one bit of information to be included in an exercise "report." An exercise "score" might say that 90 percent[11] of all 17-year-olds know the name of the president of the United States. An exercise "report" might include that information *along with* the added information that 4 percent (see Footnote 11) thought Hubert Humphrey was the president, 3 percent (see Footnote 11) thought Senator Kennedy was the president, and 3 percent (see Footnote 11) responded "I don't know."

In the area of Writing, the results of an exercise might be reported as follows: (An age 9 example; *not* a National Assessment exercise.)

Exercise 29:

Imagine that tomorrow you are going to have a substitute teacher. What is it like to have a substitute? Does school seem different? Are you happy? Or sad? Or doesn't it make much difference? Think a while and then write a short story on the way you feel about having a substitute teacher.

Ninety percent of all 9-year-olds wrote responses that were judged to be as good as or better than these examples:

Example 1:It dosent make any diffrents I think.
Example 2:I think it does not make any difference.

[11] Ficticious percentages

Fifty percent of all 9-year-olds wrote responses that were judged to be as good as or better than these examples:

Example 1:I feel very bad when I have a substitute teacher expesilie for math and reading.
Example 2:If shes meen ill be sad and if shes nice ill be happy.

Only 10 percent of all 9-year-olds wrote responses that were judged to be as good as or better than these examples:

Example 1:Most substitutes around here are worse than the orinary teachers. I just like the ordinary ones.
Example 2:I feel happy because the substitute has a turn to teach the class.

These examples were based upon an attempt to judge the meaning or thought expressed in the 9-year-old responses. They were *not* evaluated for spelling or grammatical usage. They could have been evaluated and reported independently in terms of aspects of grammar. In the actual assessment some exercises will be reported one way, some the other way, and some may be reported both ways.

In any event the reader of the final reports will be able to judge for himself whether he agrees with the "correct" or "incorrect" responses, or with the examples judged to be low or medium or high. All scoring criteria will be reported with the exercises.

These examples of exercise reports point up the fact that "scoring" of National Assessment exercises is a diverse process. Exercises, by and large, will require one of three types of scoring:

1. *Machine scoring:* Multiple-choice exercises, for which an assessee responds by filling in an oval opposite the choice that he selects can be scored and reported routinely by machine.
2. *Short answer exercises:* Short answer exercises can be scored by individuals using a key of acceptable and unacceptable answers, or by simply coding all answers into a limited subset of responses for reporting purposes. Since an important part of the reporting is to report common "errors" as well as percentages of correct responses, the scorer will be coding

more often than judging right or wrong. Individuals to do this type of job are often college-trained persons who are available for part-time work.

3. *Longer essay responses:* The scoring of longer essay responses generally requires the use of professional people who are accustomed to reading and judging essays. The most common source of personnel is English teachers. Such readers, however, must work together under the direction of trained leaders so that all readers are in fact using the same criteria for judging the responses. Reading sessions often involve bringing groups of professionals together in one location for two or three days, during which time they develop common criteria, and spend considerable time checking and cross-checking their procedures. A fairly common goal would be to ask readers to read "holistically," for general content, for general organization and communication effectiveness rather than for details. Another, separate reading might attempt to judge the more detailed grammar, usage, and mechanics level of performance.

Thus, the scoring of National Assessment exercises will be done in part by machines, in part by highly trained clerks, and in part by professional readers. Each exercise will be reported in relation to the scoring procedure or procedures used for it.

National Assessment Is A Reporting Plan

The audiences for National Assessment results are the same groups as the ones deemed important in the development of the objectives in each subject area–the subject matter specialist, the professional educator, and the informed layman. Reporting results to these various groups probably will differ primarily in the amount of detail that is included. The subject matter specialist and the professional educators are more apt to want very detailed reports, with as many different analyses as possible. The layman is more apt to want less detail and more generalization of results–what are the general conclusions, what is the big picture?

It has long been apparent that National Assessment must develop multiple reports, of differing types and natures. Detailed, voluminous reports of every exercise selected for reporting in a

given year must be prepared. These will be the basic National Assessment reports. Such reports, however, do not lend themselves to immediate or "obvious" conclusions. Someone, hopefully many different "someones," must pore over and sort out the results in a fashion that is most meaningful. Whether or not this latter step is one that National Assessment itself should attempt, or whether it should be left to the scientist, to the classroom teacher, to the school board member is a moot question. Some people feel that National Assessment should provide information only, leaving all interpretation up to the user of the results. Others feel that some attempt at interpretation is necessary, if only to posit various hypotheses that are tenable, unless one wishes to run the risk of gross misinterpretations. This issue has not been settled yet. As is often the case a middle ground may well be found.

Ultimately, regardless of who does it, evaluative reports must be developed or the potential impact for good that lies within National Assessment results will never be realized. It may well be that various professional organizations will take the initiative to look at and evaluate the results. It may well be that college and university professors will seize upon National Assessment results and prepare their own evaluative reports. It may well be that directors of curriculum and directors of instruction in school systems will take the basic reports and prepare evaluations that have relevance for their school systems. It may be that boards of education will ask their administrative staffs to prepare evaluative reports for them. It may be that state and/or federal legislators will request evaluative studies of the results by their own advisors or by consultants whom they trust. It may be that newspapers and magazines will see newsworthy stories in the results (hopefully not just attention-getting headlines).

No project such as National Assessment will have met its basic goal unless the information that it produces is, in truth, useful to these individuals who make decisions. Information such as this should be used to improve decision-making; otherwise it has little or no meaning. Developing the means for putting useful information in the hands of educational decision-makers is the goal of National Assessment reporting.

Exercises to be Reported

The plan calls for reporting approximately 40 percent of the

exercises at the end of each assessment year. Not all exercises will be reported since it is considered essential to use many of them over again in future assessments. The potential biasing of results in a future year, if all exercises given in 1969 are reported in 1969, is the problem. It is not clear whether the publication of all exercises would result in some teachers using the exercises themselves within a class, for direct instruction. However, since that possibility exists it was deemed preferable to withhold many of them until after the second and third cycles. A check on the effect of releasing exercises will be made by reusing about 10 percent of the same exercises in the second cycle and preparing about 30 percent completely new exercises to replace ones that are reported.

Reporting Categories

Various reporting categories have been and are being developed. The basic categories are the four different age groups–9, 13, 17, 26-35. In a few instances the same exercises will be used across three or even across all four age groups. In many instances the same exercise will be used at two different ages. Thus it will be possible to see some comparative data across two or more age levels. The choice of age groups to assess was made in order to sample near the end of primary education, near the end of elementary education, near the end of secondary education, and after most adults have completed all of their formal education.

A second set of reporting categories is by geographic region. Four regions were used in the sampling; the same four will be used for reporting purposes–Northeast, Southeast, Central, and West. The Northeast includes all the middle-Atlantic and New England states. The Southeast contains most border states between the North and the South plus the "deep south" and eastern Texas. The West contains all Rocky Mountain, Southwestern, and West Coast states, plus Hawaii, Alaska, and west Texas. The Central area contains the other states. These divisions correspond closely to geographic divisions used for many other statistical reports. Whether or not geographic differences appear remains to be seen.

A third set of reporting categories is based upon type of community, basically size. These categories are:

Large cities (above 200,000 population)
Urban fringes (cities adjacent to the large cities)

Middle size cities (25,000 to 200,000)
Small town-rural areas (below 25,000)

In addition to these basic reporting categories, additional information about community size and type was collected. Thus, it may be possible to report finer breakdowns by separating out central cities areas or truly rural areas, if the data warrant.

A fourth reporting category will be sex. Numerous studies have demonstrated that boys and girls produce different results in various subject areas. Such differences certainly will appear in National Assessment results.

A fifth reporting category is labeled SES. Originally it meant socio-economic status. It now is read as socio-educational status. Neither term may truly represent the intent of this breakdown. The intent was to be able to report results separately for assessees from disadvantaged homes. The great concern of contemporary society with the education of the disadvantaged requires an all-out effort to provide information about the knowledges and skills of that group as they exist today.

Defining SES or describing the intent of the classification is simple. Finding a good index or indices of SES is extraordinarily complex. The literature of educational measurement yields numerous attempts at measuring SES, each one of which has some major flaw. Ideally one might want to classify assessees according to parental income. In practice such information cannot be collected as it verges upon invasion of privacy. Obvious substitutes are educational levels of parents and/or occupational levels of parents. Such information can be secured relatively easily for 17-year-olds by direct questioning. But 9-year-olds are not apt to have the information. One can consider the use of existing school records (complete in some schools; incomplete in others) or one can consider trying to get the information directly from parents (a tedious and only partially successful scheme). Or one can ask a series of simple questions that 9-year-olds and 13-year-olds can answer, such as whether a home contains an encyclopedia or a daily newspaper or books, etc., and infer family educational level from such indices. National Assessment is trying all of these approaches in the hope that one or more of them will provide a meaningful breakdown into two or more meaningful SES levels.

The final reporting category is race. This category was added to the reporting scheme less than two years ago at the urging of

persons concerned with obtaining maximum information about minority groups. It is a controversial category that offends some people if included and offends others if omitted. The policy committee for National Assessment felt that the need for this type of information outweighed the dangers inherent in collecting it.

An additional practical reason for including race as a reporting category is the fact that it offers an additional category to be used in connection with low SES reports. Many persons assume, incorrectly, that most low SES individuals are members of a minority group. Statistics, however, say that more members of the white majority in this country are low SES than are members of minority groups in the country as a whole. The use of race as a separate reporting category will enable one to look at SES and race both independently and together.

Unfortunately, the small size of the National Assessment sample means that meaningful statistics will be available only for black, white, and other or for simply black and other. The designation of race is being made by the exercise administrators. No individual assessee is asked to indicate his race. While this is not a perfect categorization, no other scheme is perfect either. It is a categorization that is close to common usage.

A Typical Report (hypothetical data)[12]

Science: Age 17

Objective I. Know fundamental facts and principles of science.

Exercise 1. In addition to water and the correct temperature, what else do most seeds need in order to sprout?

a. light
b. soil
c. minerals
d. air (the answer)
e. I don't know.

[12]The exercise used in this illustration is *not* an actual exercise used in the assessment.

National Results:	Percentage
Light	43
Soil	19½
Minerals	9½
Air*	25
I don't know.	3

This exercise proved to be a difficult one; only one-fourth of the sample knew that air is an essential for sprouting, rather than light or soil or minerals. Light is the most common distractor, since 43 percent felt that it is essential. It could be that light and soil are perceived to be "logical" choices to someone who doesn't know the answer. Most persons associate light and soil with growth even though they are not essential for sprouting. Since only 3 percent of the assessees chose "I don't know" it would seem logical to infer that most assessees felt confident that they knew the answer, even though most did not.

Regional Results:	Northeast	Southeast	Central	West
Light	48%	44%	36%	42%
Soil	19	16	24	16
Minerals	6	8	12	14
Air*	23	28	24	26
I don't know.	4	4	4	2

The regional results show very little difference between the four geographic areas. The range from 23 percent correct in the Northeast to 28 percent correct in the Southeast is not large enough to suggest any real regional difference in this bit of information. For some reason minerals was a less attractive distractor in the Northeast than in the West, while light was least attractive in the Central area and soil more attractive. These could be random variations.

Sex Differences	Boys	Girls
Light	35%	51%
Soil	19	20
Minerals	12	7
Air*	33	17
I don't know.	1	5

The sex differences are rather pronounced. Twice as many boys as girls knew the answer. Girls were distracted most by the alternative "light," perhaps the most attractive one to a person not knowing the answer. More girls than boys chose "I don't know," but it still was a small percentage. Boys were less distracted by "light" than girls.

These results would also be broken down by type of community, by SES, and by race, and reported.

Additional breakdowns could be presented; additional exercises related to this objective could be presented. The basic report in this example attempts to "read" the results, to point up likenesses and differences. What these likenesses and differences mean would require the knowledgeable interpretation of persons thoroughly grounded in science and science education at age 17. If, for example, a whole series of exercises relating to Objective I would show sex differences, a generalization could be formulated. Only the results themselves will provide the information that can serve as a basis for such generalizations.

It should be mentioned again that no scores or reports for individuals will be made. No individual assessee took more than one-twelfth of the exercises; no individual took a package that sampled a single subject area. Therefore, individual scores were impossible (meaningless) to obtain.

Media

The basic media for National Assessment results will be the written word. However, it is anticipated that radio, television, film, personal reports, etc. will be utilized also. These other media will be particularly useful for wide dissemination of the results,

and for stimulating a wide audience of potential users of the results to seek them out and make use of them.

Using National Assessment Results

It is difficult and dangerous to speculate too much about actual uses of National Assessment results. It is anticipated that the results will be very helpful in educational decision-making. To discuss how results may be used in decision-making requires actual data, which we do not have yet. We might "make-up" some data to use as examples of how decisions might be effected. In making up results one tries to be realistic and develop hypothetical statistics that seem to make sense. However, this procedure then leads some critics to suggest that if one already "knows" what the results are going to be, why is it necessary to spend large amounts of money to simply confirm one's feelings. The answer to this, of course, is that "feelings" are not "information." At this point in time we have all sorts of opinions about what students know. What is needed is definitive information about what they do know.

Nevertheless, in spite of the problems of hypothesizing statistics let's assume that in the first year of the assessment the results in Science, in general, show that school age students exhibit quite a bit of knowledge of scientific facts, of scientific principles, and of scientific generalizations. Let's assume also that they exhibit a lower level (judged by what science teachers and others feel they should have attained) of proficiency in applying scientific principles, in actually using these principles as they solve or attempt to solve problems of everyday life. If these two bits of information should emerge from National Assessment it could well lead laymen and educators alike to take a close look at the scientific curricula, to see how greater emphasis could be placed on developing application skills in the scientific domain. This is a very general, broad example of the type of decision-making that could be influenced by information available from National Assessment. Much of the decision-making would be more specific, and would relate to specific information.

National Assessment Results Will Not Be Standards

A common, and incorrect, assumption that many users of test

results make is that the test results themselves, in some magic fashion, define what is right or good or proper.

Test results are *not* appropriate standards of achievement.

National Assessment results will *not* be appropriate standards of achievement.

Appropriate standards of achievement should be and must be determined by persons knowledgeable in a subject field and knowledgeable about the abilities that youngsters of a given age bring to the learning process. A very important ingredient in determining such standards is a knowledge of the levels of achievement at which students are functioning. But present levels of achievement are not necessarily appropriate standards themselves.

National Assessment Is A Research Project

National Assessment, as a project to secure information not currently available, can be considered as a research project itself. It does not, however, fall in the category of a project with hypotheses to be tested. It is not designed to seek relationships between the information gathered and other characteristics of schools or classrooms. These are legitimate goals, of course, but they are not goals of National Assessment at this time.

Within National Assessment it has been necessary to conduct a number of specific research studies to seek answers to questions that needed resolution in planning the project itself. Most of the specific research studies conducted for National Assessment have been summarized in a separate paper.[13] They dealt with such things as meaningfulness of the National Assessment exercises (resulting in the development of an elaborate review process of the exercises); in-school versus out-of-school administration (resulting in an assurance that results for out-of-school 17s would not be biased by the setting in which the administration took place);

[13]Womer, Frank B., "Research Toward National Assessment," *Proceedings: 1968 Western Regional Conference on Testing Problems*, Educational Testing Service: Berkeley, pp. 34-49.

modes of administration (resulting in the use of completely taped administration for group-administrable exercises); regional voice studies (resulting in the use of a male radio announcer with a "National" voice for taping the administrations); mathematics study (resulting in confirmation of the decision to use an "I don't know" alternative in multiple-choice exercises); choices study (resulting in a decision to use open-ended exercises more often than originally planned); an SES study (resulting in the development of questions relating to the presence or absence of certain "cultural-educational" items in the home as indices of SES for lower age groups).

National Assessment results, once they become generally available will most certainly raise a multitude of questions as to why the results are what they are. The project is designed primarily to provide basic information not currently available; many people will want to investigate potential reasons for the results. No doubt many research efforts will be generated to attempt to relate previously available information to the new information available from National Assessment. It may be that the research that develops as a spin-off from the questions raised by National Assessment results will prove to be as valuable to American education as the results themselves—only time will tell.

National Assessment Is An On-Going Project

The plan for National Assessment, as it exists in 1970, calls for a series of cycles, designed to provide comparable results for a given subject matter area every few years. The ultimate goal of National Assessment is the measurement of change (progress) in knowledges, skills, understandings, and attitudes as they relate to meaningful educational objectives. The only way to assess change is through repeated measurement of the same objectives with the same exercises. Thus, each subject area is to be repeated periodically in order to determine whether change does occur.

As originally conceived, each National Assessment cycle was three years in length, with three subject areas being assessed in each of two different years, and four subject areas in the third year. The first year of the first cycle is being completed on that basis, with Citizenship, Science, and Writing as the subject areas.

As National Assessment moved into its first assessment year and

as redevelopment began in the first three subject areas, it became apparent that a three-year cycle had serious flaws. If one is to use the results of one assessment to do a better job the next time in a given subject area, a longer cycle is essential. With that in mind a combination three year - six year cycle has been developed as follows:

Cycle 1

March, 1969 - February, 1970:	Science, Writing, Citizenship
October, 1970 - August, 1971:	Reading, Literature
October, 1971 - August, 1972:	Music, Social Studies
October, 1972 - August, 1973:	Math, Science, COD
October, 1973 - August, 1974:	Reading, Writing, Listening & Speaking (new)
October, 1974 - August, 1975:	Citizenship, Art, Consumer Education (new)

Cycle 2 (Oct. to Aug.)

1975 - 76	Math, Science, Health Education (new)
1976 - 77	Reading, Literature, Physical Education (new)
1977 - 78	Music, Social Studies, Study Skills (new)
1978 - 79	Math, Science, COD
1979 - 80	Reading, Writing, Listening & Speaking
1980 - 81	Citizenship, Art, Consumer Education

This cycling plan may also be modified as time goes by and additional experience is gained. Its notable features are that three areas, Reading, Mathematics, and Science, are on a three-year cycle whereas all other subject areas are on a six-year cycle. It also

shows five "new" subject areas. These are not all firm decisions at this time.

National Assessment Is A Cooperative Project

The original ad hoc committee that was called together by Francis Keppel and John Gardner soon was replaced by the Exploratory Committee on Assessing the Progress of Education (ECAPE), a non-profit corporation of the state of New York, charged with the development of a plan to assess the outcomes of education in this country and the instrumentation to implement it. ECAPE was the governing body of National Assessment from its beginning until July 1 of 1968. During that four-year period the Carnegie Corporation and the Fund For The Advancement of Education supplied all of the ECAPE funding, approximately $2 million.[14] During that period ECAPE was an 11-member committee, consisting of Ralph Tyler as chairman, with a state commissioner of education, an associate commissioner of education, a school superintendent, a high school principal, two college presidents, two businessmen, an educational consultant, and an officer of the Carnegie Corporation as members.

During that four-year exploratory period the development of the details of the assessment plan was handled primarily by a Technical Advisory Committee[15] as ECAPE concerned itself primarily with policy and general procedures. An ECAPE staff,[16] was developed to coordinate, oversee, and monitor the efforts of various agencies which were contracted with for the development of different phases of the project. Development of the objectives and the exercises for the 10 subject areas was handled by the Educational Testing Service, the American Institutes for Research, Science Research Associates, and the Psychological Corporation. The development of the sampling plan was done by the Research Triangle Institute. Other agencies that conducted special studies were the National Opinion Research Center, Eastern Regional Institute for Education, and the Southeastern Education Laboratory.

[14]The USOE granted the University of Minnesota $100,000, during the developmental period, to hold the lay conferences for review of exercises for potential offensiveness.

[15]John Tukey, Chairman; Robert Abelson; Lee Cronbach; Lyle Jones; Ralph Tyler.

[16]Directors were Stephen Withey, Jack Merwin, and Frank Womer.

Perhaps even more indicative of the cooperative nature of the entire project is the fact that over 500 separate consultants were called upon to advise ECAPE, its contractors, and its staff in the many different details of the project. Some of these consultants advised the contractors in the development of the objectives; others advised staff in its reviews of the objectives. Considerable use of consultants was made by staff in the very extensive process of review of exercises. Many of these consultants were subject matter specialists, others were lay persons interested in and knowledgeable about education.

There are few aspects of National Assessment that could be traced to a single source; most of the plan and instrumentation as it exists in 1970 is a cooperative plan that reflects the thinking of many persons. National Assessment is truly a cooperative project.

On July 1, 1968, ECAPE became CAPE by dropping the term "exploratory" and moving into the operational phase of the project. Since then two grants from the U.S. Office of Education, a grant from the Carnegie Corporation, and a grant from the Ford Foundation have been received, totaling close to $3 million. This was sufficient funding to carry the project through most of 1969. Funding for fiscal 1970 is primarily from the Office of Education.

The change from ECAPE to CAPE was accompanied by an expansion of the Committee from 11 to 23 members. This expansion permitted the representation of individuals associated with an even wider group of organizations and agencies working in and allied to education, such as the American Association of School Administrators, the Chief State School Officers, the National Association of Secondary School Principals, the Department of Elementary School Principals, the National Education Association, the American Federation of Teachers, the National Congress of Parents and Teachers, the National Association of State Boards of Education, the National School Boards Association, and so on.

A concern expressed by some with the governance of the project was the self-perpetuating aspect of CAPE, the fact that CAPE members were not responsible, directly or indirectly, to the electorate. In consideration of that concern the Education Commission of the States was approached and asked to consider becoming the governing body for National Assessment. In June of 1969 the Steering Committee of ECS voted to accept governance of the project. On July 1, 1969, National Assessment became one

of the projects under the general supervision of the Education Commission of the States.

National Assessment Is A Changing Project

All of the foregoing statements about National Assessment, about its various policies and plans, are subject to continuing review. None is so firm that it may not be altered or even reversed if a new policy seems to make sense or if an altered practice gives promise of more effective assessment. The policies and plans as they exist satisfy a lot of people. They are too ambitious for some; they fall short of hoped-for goals of others. When a review of the project is written a year from now, it is almost certain that some of the policies and plans outlined in this paper will have changed.

There are a number of areas within the project that are subject to considerable discussion, primarily because some people prefer an alternate approach. A few of these will be mentioned here.

Should National Assessment attempt to relate its results to other educational statistics? The present plan does not include gathering information about sample schools, information such as per pupil expenditures, curricula, staff characteristics, etc., the sorts of things that often are hypothesized to be directly related to educational outcomes. This type of addition to the project could be handled fairly easily from an administrative point of view. It would require some additional effort on the part of cooperating schools.

Is National Assessment truly sampling the objectives within each subject area? Budgetary limitations have placed a ceiling of about 160 to 180 minutes of assessment time that is available for each subject area for each age level. The question then arises as to whether the area of Science for age 9 assessees can adequately be sampled in 170 minutes, or whether the area of Citizenship for age 13 can adequately be sampled in 160 minutes, or whether the area of Writing for age 17 can adequately be sampled in 180 minutes. The answer probably is, "No, but it's more time than has ever been devoted before to an assessment of knowledges and skills in a subject matter area for a given age level." Any increase in coverage would increase costs of the project significantly.

Should National Assessment be assessing age 5 youngsters, in order to provide information about knowledges and skills prior to entry into formal education? The original thinking on this matter

was that it would be wise to assess after the educational system had had an opportunity to have an impact on students. Yet, "before" information could be very useful. Again, budget would have to be expanded significantly to consider this alternative–the administrative portion by at least one-fourth, plus all the developmental work that would be necessary.

Should National Assessment remain "National"? The plan, as developed over the last five years, was sensitive to the concerns of many critics, particularly school administrators, that the real goal of the project was to evaluate individual school districts or individual states. The sampling plan, set up by fairly large geographic regions, does not permit such comparisons. Yet, in the last year more and more concerns are being expressed that, since results will not be available for states or school districts, one level of potential usefulness has been omitted. It will be possible, of course, for states and/or school districts to use the same exercises used in National Assessment, *after* the National Assessment results are published. Whether or not any more direct ties should be developed is a question that undoubtedly will be debated seriously in the next several years. Any expansion of the present project would be dependent upon budget, as usual.

Should new areas be added to National Assessment? A part of the original plan for National Assessment was that 10 areas were as many as could reasonably be undertaken at the beginning. It was never felt that the 10 areas selected covered all of the important areas. In order to develop a new subject area, a minimum of three or more years is essential (it took four years to develop the first three areas). Thus nothing new could be introduced into the assessment prior to 1973 (if the process began by 1970). Some thought has already been given to new areas, and specifications have been developed for four of them–Listening and Speaking, Health Education, Consumer Education, and Physical Education. At this point no work is being done on new areas.

Will National Assessment raw data be available to other researchers? The present policy is "no, not now." National Assessment is a new and bold educational project. If it is to survive it must concentrate every effort toward its own primary goal of gathering and reporting important information not now available to the educational community. It must prove itself before it can even consider the vast potentialities of making its data available to others; it must do its own thing first. Once National Assessment is

accepted as an on-going project, it then can and must consider how best to use the masses of data that it has collected. Undoubtedly one way to maximize its use will be to make it available to educational researchers who have questions which National Assessment data can answer.

What Is National Assessment?

National Assessment is an educational project. National Assessment is an educational project designed to provide information not currently available about many of the direct outcomes of education, knowledges, skills, understandings, and attitudes, in a variety of subject areas. The purpose behind National Assessment is to improve the educational decision-making of legislators, board of education members, professional educators, and all others vitally concerned with improving American education. It is assumed that decision-making is improved by providing information pertinent to the decision-making.

Progress in education is dependent upon knowing where we stand at a given time, how far we have come, and the direction in which we are going. National Assessment has the potential of focusing national attention upon our present status, our past performance, and the direction in which we are moving.